La Inteligencia Artificial y su costo energético:

Uso actual, límites del escalamiento y alternativas estructurales

Introducción

El consumo energético asociado a la Inteligencia Artificial se ha convertido en un tema de creciente interés y preocupación. Lo que durante años fue considerado un costo técnico asumible hoy aparece como un factor estructural que condiciona la escalabilidad, la sostenibilidad y el diseño futuro de los sistemas de IA.

Este documento analiza el problema energético de la IA desde una perspectiva temporal —inicio, presente y mediano plazo— y desde una mirada funcional: cómo y por qué se consume energía. A partir de este análisis, se propone que parte del problema no reside únicamente en la infraestructura o las fuentes energéticas, sino en la **estructura cognitiva de los modelos** y en la **calidad de la interacción con los usuarios**. Finalmente, se presenta la **Teoría G** como una vía investigable orientada a reducir consumo innecesario mediante mejoras estructurales.

1. Al inicio: cuando la energía no parecía un problema

En las primeras fases del desarrollo de la IA moderna, el consumo energético no ocupaba un lugar central en el debate. Los modelos eran más pequeños, los casos de uso estaban acotados y el entrenamiento ocurría de forma relativamente esporádica.

El progreso se midió principalmente en términos de precisión, rendimiento y escalamiento. El aumento del consumo energético se interpretó como un efecto colateral inevitable del avance tecnológico, bajo la suposición implícita de que la eficiencia sería resuelta por el hardware y las mejoras en infraestructura.

En ese contexto, el paradigma dominante fue claro: **más datos y más cómputo producirían sistemas cada vez más inteligentes**.

2. La situación actual: el problema se vuelve visible

Hoy el escenario es distinto. La IA no solo se entrena; se **utiliza de forma continua y masiva**. Millones de usuarios interactúan diariamente con modelos que ejecutan inferencias de manera permanente, muchas veces para tareas de bajo valor informativo.

Un punto crítico es que los modelos actuales **no distinguen entre una consulta bien formulada y una consulta deficiente**. Desde el sistema, ambas activan procesos de inferencia comparables. Esto implica que una parte relevante del consumo energético diario se destina a generar respuestas plausibles para preguntas ambiguas, redundantes o mal planteadas.

El problema, por tanto, ya no es solo cuánto consume cada operación, sino **cuántas operaciones se ejecutan sin necesidad real**.

3. Rangos técnicos comparativos de consumo energético

Las cifras que siguen son **rangos aproximados**, útiles para comprender órdenes de magnitud. No buscan exactitud absoluta, sino contextualización comprensible.

Tabla comparativa de consumo energético estimado:

Actividad	Consumo aproximado	Equivalencia cotidiana
Consulta simple a un modelo de lenguaje	0.002 – 0.01 kWh	Bombilla LED de 10 W encendida 10–60 minutos
Consulta compleja (texto largo, razonamiento)	0.02 – 0.05 kWh	Laptop en uso continuo 2–4 horas
50–100 consultas en un día	0.3 – 1 kWh	Refrigerador doméstico funcionando un día
200–300 consultas en un día	1 – 2 kWh	Aire acondicionado doméstico 1–2 horas
Miles de usuarios simultáneos	Decenas de MWh/día	Edificio corporativo completo en operación
Entrenamiento de un modelo grande	1 – 10 GWh	Cientos o miles de hogares durante semanas

Para referencia adicional:

■ Un aire acondicionado doméstico consume entre **0.8 y 1.5 kWh por hora**.
■ Un hogar promedio consume entre **100 y 200 kWh al mes**.
■ Un centro de datos de IA puede alcanzar consumos equivalentes a **una ciudad pequeña**.

Estas comparaciones muestran que el impacto energético de la IA **no es trivial** cuando se observa de forma agregada.

4. Qué están haciendo actualmente las empresas tecnológicas

Las empresas líderes han respondido principalmente desde la infraestructura:

■ Desarrollo de hardware especializado (GPUs, TPUs, aceleradores).

■ Optimización térmica y eficiencia en centros de datos.

■ Migración hacia energías renovables y contratos energéticos dedicados.

■ Técnicas de optimización de modelos: compresión, cuantización, fine-tuning.

Estas medidas son necesarias y efectivas, pero comparten una característica común: **reducen el costo por inferencia**, no necesariamente **el número de inferencias**.

5. El punto ciego del paradigma actual

Desde el punto de vista funcional, los modelos actuales están diseñados para **responder siempre**. No poseen mecanismos explícitos para:

■ Evaluar si una pregunta está bien definida.

■ Detectar inconsistencias conceptuales tempranas.

■ Decidir que el costo computacional no justifica la inferencia.

■ Pedir reformulación antes de ejecutar procesos costosos.

Desde una perspectiva energética, esto implica procesar información incluso cuando **no hay un problema válido que resolver**.

6. Mediano plazo: qué es razonable esperar

Si el paradigma actual se mantiene sin cambios estructurales, es razonable anticipar:

■ Incremento sostenido del consumo energético global asociado a IA.

■ Traslado parcial de costos al usuario final.

■ Mayor presión regulatoria y social.

■ Límites prácticos al escalamiento continuo.

Esto sugiere que la sostenibilidad futura de la IA no depende únicamente de nuevas fuentes de energía, sino de **usar el cómputo de forma más inteligente**.

7. Una vía complementaria: mejorar estructura antes que escalar

Una estrategia adicional consiste en **reducir inferencias innecesarias**, no solo hacerlas más eficientes. Esto requiere modelos capaces de:

■ Evaluar la calidad estructural de la consulta.

■ Pedir aclaraciones o reformulaciones cuando sea necesario.

■ Incorporar decisión negativa: no responder cuando el problema no está definido.

■ Guiar al usuario hacia mejores formulaciones.

Incluso una reducción del **10–20 %** de inferencias innecesarias, a escala global, tendría un impacto energético comparable o superior a muchas mejoras marginales de hardware.

8. La Teoría G como opción investigable

En este marco se inscribe la **Teoría G**, una línea teórica en desarrollo que propone complementar la IA probabilística con estructura explícita de análisis, decisión y límites.

Al introducir conceptos como:

■ Invariantes cognitivos en la interacción usuario–sistema.

■ Evaluación estructural del problema antes de inferir.

■ Capacidad explícita de decir “no puedo” o “no debo”.

La Teoría G apunta a **usar mejor el cómputo**, no a incrementarlo.
Los postulados iniciales de esta línea han sido registrados públicamente en Archive.org como documentos de trabajo abiertos a evaluación y desarrollo.

Conclusión

El consumo energético de la Inteligencia Artificial ya no es un efecto colateral menor, sino una variable estructural que condiciona su evolución. Resolver este desafío requiere combinar avances en infraestructura con una revisión profunda de cómo los sistemas deciden cuándo y cómo responder.

Reducir inferencias innecesarias, mejorar la calidad de la interacción y dotar a los modelos de criterios explícitos de decisión puede ser tan relevante para la sostenibilidad como cualquier innovación energética. En este contexto, enfoques teóricos como la Teoría G ofrecen una vía complementaria que merece ser explorada con rigor y prudencia.